Student Alcohol Consumption Prediction: Data Mining Approach

[1]Hind Almayyan, [2]Waheeda Almayyan

[1]Computer Department, Institute of Sectary Studies, PAAET, Kuwait
hi.almayyan@paaet.edu.kw

[2]Computer Information Department, Collage of Business Studies, PAAET, Kuwait
wi.almayyan@paaet.edu.kw

Abstract

Alcohol consumption in higher education institutes is not a new problem; but excessive drinking by underage students is a serious health concern. Excessive drinking among students is associated with a number of life-threatening consequences that include serious injuries; alcohol poisoning; temporary loss of consciousness; academic failure; violence, unplanned pregnancy; sexually transmitted diseases, troubles with authorities, property damage; and vocational and criminal consequences that could jeopardize future job prospects. This article describes a learning technique to improve the efficiency of academic performance in the educational institutions for students who consume alcohol. This move can help in identifying the students who need special advising or counselling to understand the danger of consuming alcohol. This was carried out in two major phases: feature selection which aims at constructing diverse feature selection algorithms such as Gain Ratio attribute evaluation, Correlation based Feature Selection, Symmetrical Uncertainty and Particle Swarm Optimization Algorithms. Afterwards, a subset of features is chosen for the classification phase. Next, several machine-learning classification methods are chosen to estimate the teenager's alcohol addiction possibility. Experimental results demonstrated that the proposed approach could improve the accuracy performance and achieve promising results with a limited number of features.

Keywords

Data mining; Data mining; Classification; Student's performance; Feature selection; Particle swarm optimization; Alcohol consumption prediction.

1. INTRODUCTION

Globally, heavy alcohol drinking is associated with premature death, weaker probability of employment, more absence from work, in addition to lost productivity and lower wages. Moreover, alcohol consumption results in approximately 3.3 million deaths each year [1]. It is the third largest risk factor for alcohol-related hospitalizations, deaths and disability in the world. Approximately one in four children younger than 18 years old in the United States is exposed to alcohol abuse or alcohol dependence in the family [2]. Alcohol consumption has consequences for the health and well-being of those who drink and, by extension the lives of those around them.

The relationship between problematic alcohol consumption and academic performance is a concern for decision makers in education. [3] Alcohol consumption has been negatively associated with poor academic performance, [4] and heavy drinking has been proposed as a probable contributor to student attrition from school. [5]

Traditional methods for monitoring adolescent alcohol consumption are based on surveys, which have many limitations and are difficult to scale. Therefore, several approaches have been investigated using conventional and artificial intelligence techniques in order to evaluate the teenage alcohol consumption. In Crutzen et al. [6] a group of Dutch researchers studied the association between parental reports, teenager perception and parenting practices to identify binge drinkers. They designed a binary classifier using alternating decision trees to establish the effectiveness of the results of exploring nonlinear relationships of data. Montaño et al. [7] proposed an analysis of psychosocial and personality variables about nicotine consumption in teenagers. They applied several classification techniques such as RNA Multi-layer perceptron, radial basis functions and probabilistic networks, decision trees, logistic regression model and discriminant analysis. They discriminated successfully 78.20% of the subjects, which indicates that this approach can be used to predict and prevent similar addictive behavior.

Pang et al. [8] applies a multimodal study to identify alcohol consumption in an audience of minors, specifically the users of the Instagram social network. The analysis is based on facial recognition of selfie photos and exploring the tags assigned to each image with the objective of finding consumption patterns in terms of time, frequency and location. In the same way, they measured the penetration of alcohol brands to establish their influence in the consumption behavior of their followers. Experimental results were satisfactory and compliant with the polls made in the same audience, which can lead to use this approach to other domains of public health.

In Bi et al. [9], a study using two machine learning methods to identify effectively the daily dynamic alcohol consumption and the risk factors associated to it. For this, they proposed a Support Vector Machine (SVM) as classifier to establish a function for stress, state of mind and consumption expectancy, differentiating drinking patterns. After that, a fusion between clustering analysis and feature classification was made to identify consumption patterns based on daily behavior of average intake and detect risk factors associated to each pattern. Zuba et al. [10] proposed machine learning approach that use a feature selection method with 1-norm support vector machines (SVM) to help classify college students between high risk and low risk alcohol drinkers and the risk factors associated to the heavy drinkers. This approach could be used to help to detect early signs of addiction and dependence to alcohol in students.

In this article, we are addressing the prediction of teenager’s alcohol addiction by using past school records, demographic, family and other data related to student. This article extends the research conducted by Cortez and Silvain in 2008 [11]. This study seeks to establish the correlation between poor academic performance and the use of alcohols among teenagers. We applied several data mining tools and ends of evaluation shows potential of better results. This article suggests a new classification technique that enhances the student performance prediction using less number of attributes then the ones used in the original research. The aim is getting better prediction results using less parameters in the process.

The article starts the suggested approach is presented in Section 3. Section 4 describes the experiment steps and the involved dataset. Section 5 shows the experiment result. The article concluded with conclusion and further research plan.

2. The Proposed Approach

Initially, several machine-learning classification methods, which are considered very robust in solving non-linear problems, are chosen to estimate the class possibility. These methods include feed-forward artificial Neural Network with MLP, Simple Logistic multinomial logistic model, Rotation Forest, Random Forest ensemble learning methods and C4.5 decision tree and Fuzzy Unordered Rule Induction Algorithm (FURIA) classifiers. We carried out extensive experimentation to prove the worth of the proposed approach. We analyze the results of the dataset from each of the perspectives of, Accuracy, ROC and Cohen's kappa coefficient. Feature extraction has played a significant role in many classification systems [12]. On this basis, the focus of this section is on the applied feature selection techniques.

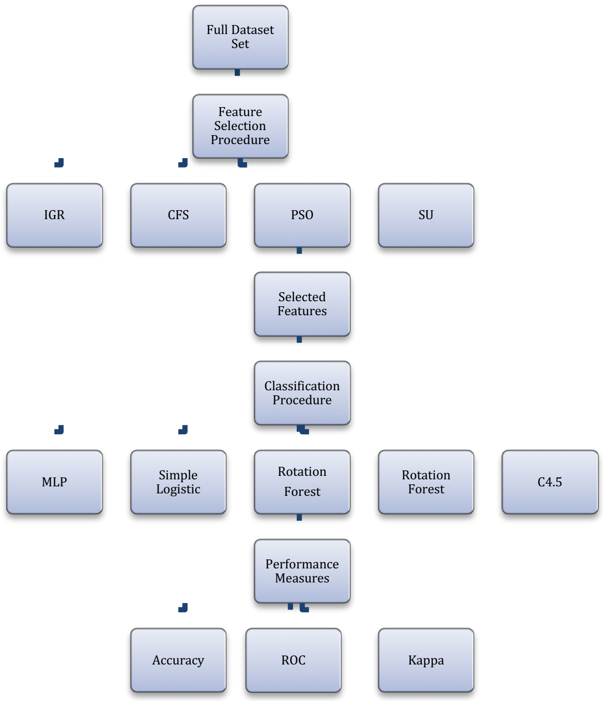

Figure 1 The proposed methodology

2.1 Particle swarm optimization (PSO)

The PSO technique is a population-based stochastic optimization technique first introduced in 1995 by Kennedy and Eberhart [13]. In PSO, a possible candidate solution is encoded as a finite-length string called a particle p_i in the search space. All of the particles make use of its own memory and knowledge gained by the swarm as a whole to find the best solution. With the purpose of discovering the optimal solution, each particle adjusts its searching direction according to two features, its own best previous experience (p_{best}) and the best experience of its companions flying experience (g_{best}). Each particle is moving around the n-dimensional search space S with objective function $f : S \subseteq \Re^n \rightarrow \Re$. Each particle has a position $x_{i,t}$(t represents the iteration counter), a fitness function $f(x_{i,t})$ and ''flies'' through the problem space with a velocity $v_{i,t}$. A new position $z_1 \in S$ is called better than $z_2 \in S$ iff $f(z_1) < f(z_2)$.

Particles evolve simultaneously based on knowledge shared with neighbouring particles; they make use of their own memory and knowledge gained by the swarm as a whole to find the best solution. The best search space position particle i has visited until iteration *t* is its previous experience p_{best}. To each particle, a subset of all particles is assigned as its neighbourhood. The best previous experience of all neighbours of particle i is called g_{best}. Each particle additionally keeps a fraction of its old velocity. The particle updates its velocity and position with the following equation in continuous PSO [14]:

$$v_{pd}^{new} = \omega * v_{pd}^{old} + C_1 * rand_1() * (pbest_{pd} - x_{pd}^{old}) + C_2 * rand_2() * (gbest_{d_d} - x_{pd}^{old}) \qquad 1$$

$$x_{pd}^{new} = x_{pd}^{old} + v_{pd}^{new} \qquad 2$$

The first part in Equation 1 represents the previous flying velocity of the particle. While the second part represents the ''*cognition*'' part, which is the private thinking of the particle itself, where C_1 is the individual factor. The third part of the equation is the ''*social*'' part, which represents the collaboration amongst the particles, where C_2 is the societal factor. The acceleration coefficients (C_1) and (C_2) are constants represent the weighting of the stochastic acceleration terms that pull each particle toward the p_{best} and g_{best} positions. Therefore, the adjustment of these acceleration coefficients changes the amount of 'tension' in the system. In the original algorithm, the value of ($C_1 + C_2$) is usually limited to 4 [14]. Particles' velocities are restricted to a maximum velocity, V_{max}. If V_{max} is too small, particles in this case could become trapped in local optima. In contrast, if V_{max} is too high particles might fly past fine solutions. According to Equation 1, the particle's new velocity is calculated according to its previous velocity and the distances of its current position from its own best experience and the group's best experience. Afterwards, the particle flies toward a new position according to Equation 2. The performance of each particle is measured according to a pre-defined fitness function.

2.2 Information Gain Ratio (IGR) attribute evaluation

IGR measure was generally developed by Quinlan [15] within the C4.5 algorithm and based on the Shannon entropy to select the test attribute at each node of the decision tree. It represents how precisely the attributes predict the classes of the test dataset in order to use the 'best' attribute as the root of the decision tree.

The expected IGR needed to classify a given sample s from a set of data samples C IRG(s,C) is calculated as follow

$$IGR(s,C) = \frac{gain(s,C)}{split\_\inf o(C)}, \qquad 4$$

$$gain(s,C) = entropy(s,C) - entropy_p(s,C),$$

$$entropy(s,C) = -p(s|C)\log_2 p(s|C) - (1 - p(s|C))\log_2(1 - p(s|C)),$$

$$p(s,C) = freq(s|C) / |C|,$$

$$entropy_p(s,C) = \sum_i \frac{|C_i|}{|C|} entropy_p(s,C_i),$$

$$split\_\inf o(C) = -\sum_i \frac{|C_i|}{|C|} \log \frac{|C_i|}{|C|},$$

where freq(s,C), C_i and $|C_i|$ are the frequency of the sample s in C, the ith class of C and the number of samples in C_i, respectively.

2.3 Symmetrical Uncertainty

Symmetric uncertainty correlation-based measure (SU) can be used to evaluate the goodness of features by calculating between feature and the target class [16]. The features having greater SU value gets higher importance. SU is defined as

$$SU(X,Y) = \frac{2IG(X|Y)}{(H(X) + H(Y)}, \qquad 5$$

$$IG(X|Y) = \frac{H(X)}{H(X|Y)}$$

Where H(X), H(Y) , H(X|Y), IG are the entropy of a of X, entropy of a of Y and the entropy of a of posterior probability X given Y and information gain, respectively.

2.4 Genetic Algorithms (GAs)

The basic idea behind the evolutionary algorithms (EAs) is derived from theory of biological evolution developed by Charles Darwin and others. It has been used as computational models and as adaptive search strategies for solving optimization problems. Genetic algorithms were developed in 1975 by Holland as a class of EAs [17]. GAs include a rapidly evolving population of artificial organisms, or so-called agents. Every agent is comprised of a genotype, often called a binary string or chromosome, which encodes a solution to the problem at hand and a phenotype that is the solution. In GAs, at the start the population of agents is randomly generated representing candidate solutions to the problem.

The GAs implementation relies on the appropriate formulation of the fitness function. The main objective of the closed identification fitness function is to maximize the recognition rate. Every agent is evaluated in each iteration, to produce new candidate solutions new fitter offspring and to replace weaker members of the last generation. Thus, the core of this class of evolutionary algorithms lies in selectively breeding new genetic structures along the course of finding solutions for the problem at hand [18]. We have adopted the algorithm described by Goldberg [19]. The flowchart of GA-based feature selection is described in the Figure 2 below [20].

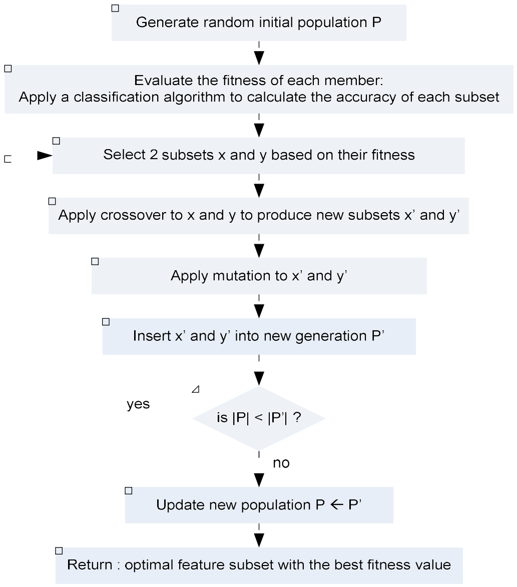
Generate random initial population P
Evaluate the fitness of each member:
Apply a classification algorithm to calculate the accuracy of each subset
Select 2 subsets x and y based on their fitness
Apply crossover to x and y to produce new subsets x' and y'
Apply mutation to x' and y'
Insert x' and y' into new generation P'
yes
is |P| < |P'| ?
no
Update new population P ← P'
Return : optimal feature subset with the best fitness value

Figure 2 Feature Selection using GA [20]

2.5 Simple random sampling

Usually real-time databases experience class imbalance problems, due to the fact that one class is represented by a considerably larger number of instances than other classes. Subsequently, classification algorithms tend to ignore the minority classes. Simple random sampling has been advised as a good means of increasing the sensitivity of the classifier to the minority class by scaling the class distribution. An empirical study where the authors used twenty datasets from UCI repository has showed quantitatively that classifier accuracy might be increased with a progressive sampling algorithm [21]. Weiss and Provost deployed decision trees to evaluate classification performances with the use of a sampling strategy. Another important study used sampling to scale the class distribution and mainly focus on biomedical datasets [22]. The authors measure the effect of the suggested sampling strategy by the use of nearest neighbor and decision tree classifiers. In Simple random sampling, a sample is randomly selected from the population so that the obtained sample is representative of the population. Therefore, this technique provides an unbiased sample from the original data.

Regarding simple random sampling there are two approaches while making random selection, in the first approach the samples are selected with replacement where the sample can be selected more than once repeatedly with an equal selection chance. In the other approach the selection of samples is done without replacement where the sample can be selected only once, so that each sample in the data set has an equal chance of being selected and once selected it cannot be chosen again [23].

3. Dataset and Evaluation Procedure

3.1 Dataset

The dataset used in this research was collected by customized questionnaire and school reports during the 2005-2006 academic year from two public schools in the Alentejo region of Portugal [11]. The school reports included few attributes such as the three period grades and number of school absences. Researchers have designed a questionnaire with closed questions to extract further socio-demographic information that were expected to affect student performance. Such information includes demographic data (e.g. mother's education, family income), social-emotional (e.g. alcohol consumption) (Pritchard and Wilson 2003) and academic learning attributes (e.g. number of past class failures) that were expected to affect student performance. The questionnaire was first reviewed by school professionals and tested on a small set of 15 students in order to get a feedback. Eventually 788 students completed the customized questionnaire. The dataset has 33 attributes, variables, or features for each student. The academic status or final student performance, which has two possible values: Pass (G3 $\geq$10) or Fail. Eventually, to find alcohol consumption, there are two different attributes related to alcohol, alcohol taking in work day (D_alc) and alcohol taking in weekend(W_alc). Therefore, the total alcohol consumption by a specific student in a whole week was estimated using the following formula [24]

$$\text{Alcohol consumption} = (W\_alc\times 2+D\_alc\times 5)/7 \qquad 6$$

The new attribute varies between one and five. Therefore, the dataset is divided into two classes according to its alcohol consumption column, which is set to 1 for the alcohol consumption is greater than 3 and 0 otherwise. The 30 features along with description are listed in Table 1.

Table 1: The dataset description of attributes [11]

Attribute Number	Attribute Description	Attribute type	Possible values of attributes
1	School - student's school	Binary	"GP" - Gabriel Pereira or "MS" - Mousinho da Silveira
2	Gender - student's gender	Binary	"F" - female or "M" - male
3	Age - student's age	Numeric	from 15 to 22
4	Address - student's home address type	Binary	"U" - urban or "R" - rural)
5	Famsize - family size	Binary	"LE3" - less or equal to 3 or "GT3" - greater than 3
6	Pstatus - parent's cohabitation status	Binary	"T" - living together or "A" - apart
7	Medu - mother's education	Numeric	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
8	Fedu - father's education	Numeric	0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
9	Mjob - mother's job	Nominal	"teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other"

10	Fjob - father's job	Nominal	"teacher", "health" care related, civil "services" (e.g. administrative or police), "at home" or "other")
11	Reason - reason to choose this school	Nominal	close to "home", school "reputation", "course" preference or "other")
12	Guardian - student's guardian	Nominal	"mother", "father" or "other"
13	Traveltime - home to school travel time	Numeric	1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
14	Studytime - weekly study time	Numeric	1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
15	Failures - number of past class failures	Numeric	(numeric: n if 1<=n<3, else 4)
16	Schoolsup - extra educational support	Binary	yes or no
17	Famsup - family educational support	Binary	yes or no
18	Paid - extra paid classes within the course subject	Binary	yes or no
19	Activities - extra-curricular activities	Binary	yes or no
20	Nursery - attended nursery school	Binary	yes or no
21	Higher - wants to take higher education	Binary	yes or no
22	Internet - Internet access at home	Binary	yes or no
23	Romantic - with a romantic relationship	Binary	yes or no
24	Famrel - quality of family relationships	Numeric	from 1 - very bad to 5 – excellent
25	Freetime - free time after school	Numeric	from 1 - very low to 5 - very high
26	Goout - going out with friends	Numeric	from 1 - very low to 5 - very high
27	Health - current health status	Numeric	from 1 - very bad to 5 - very good
28	Absences - number of school absences	Numeric	from 0 to 93
29	G3 - final grade	Numeric	from 0 to 20, output target
30	Alcohol consumption – Target class	Binary	1= yes or 0= no

3.2 Performance Analysis

The performance of the suggested technique was evaluated by using three thresholds and rank performance metrics, Accuracy, ROC and Cohen's kappa coefficient. The main formulations are defined in Equations 6-8, according to the confusion matrix, which is shown in Table 2. In the confusion matrix of a two-class problem, TP is the number of true positives that represent in our case the Pass cases that that was classified correctly. FN is the number of false negatives that represents the Pass cases that was classified incorrectly as Fail. TN is the number of true negatives, which represents the Fail cases that was classified as Fail. FP is the number of false positives that represents the Pass cases that was classified as Passed.

Table 2: The confusion matrix

Hypothesis	**Predicted patient state**	
	Classified Pass	Classified Fail
Hypothesis positive **Pass**	True Positive TP	False Negative FN
Hypothesis negative **Fail**	False Positive FP	True Negative TN

Consequently, we can define Precision as:

$$\text{Precision}= \frac{\text{TN}}{\text{FP+TN}} \times 100\% \qquad 6$$

Precision measures how many of the points predicted as significant are in fact significant. Receiver Operator Characteristic (ROC) curve is another commonly used measure to evaluate two-class decision problems in Machine Learning. The ROC curve is a standard tool for summarizing classifier performance over a range of trade-offs between TP and FP error rates [25]. ROC usually takes values between 0.5 for random drawing and 1.0 for perfect classifier performance.

Kappa error or Cohen's kappa statistics is another recommended measure to compare the performances of different classifiers and henceforth the quality of selected features. Generally, Kappa error value ∈ [-1,1], so when Kappa error value calculated for classifiers approaches to 1, then the performance of classier is assumed to be more realistic [26]. The Kappa error measure can be calculated using the following formula:

$$\text{Kappa error} = \frac{P(A)-P(E)}{1-P(E)} \qquad 7$$

where *P(A)* is total agreement probability and *P(E)* is the hypothetical probability of chance agreement.

In order to get reliable estimates for classification accuracy on each classification task, every experiment has been performed using 10-fold cross-validation. Cross-validation is a method designed for estimating the generalization error based on "resampling" [27]. Cross-validation technique allows using the whole dataset for training and testing. In k-fold cross-validation procedure, the relevant dataset is partitioned randomly into approximately equal size k parts called folds and trained k times, each time leaving out one of the folds from training process, whilst using only the omitted fold to compute error criterion. Then the average error across all k trials is estimated as the mean error rate and defined as:

$$E = \frac{1}{k}\sum_{i=1}^{k} e_i \qquad 8$$

where, e_i is error rate of each k experiment. Figure 3 depicts the concept behind k-fold cross validation.

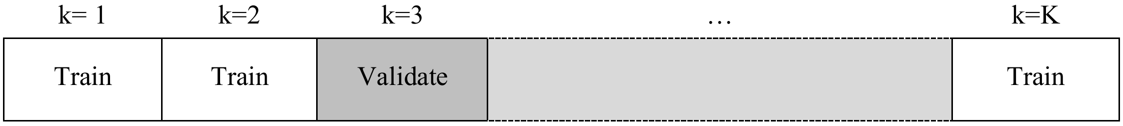

Figure 3 Data partitioning using k-fold cross-validation.

The whole dataset is divided into K folds. One-fold (k=3, in this example) is set aside to validate the data of testing and the remaining K-1 folds are used for training. The entire procedure is repeated for each of the K folds. A number of studies found that the value of 10 for k leads to adequate and accurate classification results [28].

4. Results and discussion

A multi-class classification problem such as predicting student alcohol consumption is a challenging application of data mining. The basic idea of data mining is to extract hidden knowledge using data mining techniques. The suggested system for the purpose of predicting student alcohol consumption applied in this study is carried out in three major phases. The process starts with applying the simple random sampling to scale the imbalanced class distribution. In the second phase, the feature space is searched to reduce the feature numbers and prepare the conditions for the next step. This task is carried out using four feature reduction techniques, namely GR, CFS, SU and PSO Algorithms. At the end of this step a subset of features is chosen for the next round. Afterwards, the selected features are used as the inputs to the classifiers. Five classifiers are proposed to estimate the success possibility as mentioned previously, these methods include MLP, Simple Logistic, Rotation Forest, Random Forest, C4.5 decision tree and SVM.

All the experiments were carried in Waikato Environment for Knowledge Analysis (Weka) a popular suite of data mining algorithms written in Java. The RF algorithm ensemble classifier is designed based on 150 trees and 10 random features to build each tree. While C4.5 classifier was applied with a confidence factor for pruning = 0.25 and a minimum number of instances per leaf of 2. The suggested algorithm is trained using 10-fold cross validation strategy to evaluate the classification accuracy on the dataset. Whereas the PSO feature selection was applied with a population size of 20, number of iterations = 20, individual weight = 0.34 and inertia weight = 0.33.

Table 3 depicts the effect of the class distribution before and after applying the simple random sampling technique. The unbalanced distribution of the two classes makes this dataset suitable to test the effect of simple random sampling strategy. We, therefore, used a simple random sampling approach without replacement to rescale class distribution of the dataset.

The experimental results of the multiple classifiers before and after applying the SRS can be shown in Table 4. The best overall performance is associated with Random Forest classifier with a precision of 92.2%, ROC of 94.5% and Kappa value of 70.4%, and a precision of 98.5%, ROC of 99.7% and Kappa value of 95.2% with all features before and after applying the SRS strategy with all features respectively. As for the classifiers that are used to perform predictions based on the extracted features, we observed that there is no significant difference in performance that explains the importance of SRS step.

Table 3

Class distribution of the Student dataset before and after SRS

Index	**Class**	**Class Distribution**	
		Before SRS	After SRS
1	Alcoholic	188	411
2	Not Alcoholic	856	1677

Table 4

Performance measures of selected classifiers before feature selection

Classifier	Performance index	Before SRS	After SRS
MLP	Accuracy	0.846	0.929
	ROC	0.829	0.776
	Kappa	0.479	0.776
Simple Logistic	Accuracy	0.830	0.861
	ROC	0.828	0.874
	Kappa	0.370	0.523
Random Forest	Accuracy	0.922	0.985
	ROC	0.945	0.997
	Kappa	0.702	0.952
C4.5	Accuracy	0.828	0.946
	ROC	0.732	0.933
	Kappa	0.412	0.829
FURIA	Accuracy	0.868	0.967
	ROC	0.824	0.967
	Kappa	0.476	0.885

The second phase involves searching feature vector to reduce the feature numbers and prepare the conditions for the next step. This task is carried out using four feature selection techniques, GR, CFS, SU and PSO Algorithms. The optimal features of these techniques are summarized in Table 5. As noted from Table 3, the dimensionality of features is noticeably reduced. It is worth noting that the number of features has remarkably reduced, therefore less storage space is required for the execution of the classification algorithms. This step helped in reducing the size of dataset to only 6 to 15 attributes.

Table 5

Selected features of the student dataset

FS technique	Number of selected features	Selected features
IGR	13	2,10,11,13,14,15,17,21,25,26,27,28,29
SU	15	2,5,10,11,13,14,15,17,20,21,25,26,27,28,29
GA	7	2,13,25,26,27,28,29
PSO	6	2,13,26,27,28,29

The experimental results of the multiple classifiers with the reduced number of features can be shown in Table 6. The highest performance rate for the IGR-based feature selection technique is associated with Random Forest classifier with 97.9%, 98.7% and 93.3% for Accuracy, ROC and Kappa, respectively with 13 features. While the highest performance rate for the SU-based feature selection technique is associated with C4.5 classifier with 99.5%, 99.7% and 98.2% for Accuracy, ROC and Kappa, respectively with 15 features. The highest performance rate for the GA-based feature selection technique is associated with Random Forest classifier with 96.2%, 97.2% and 88.1% for Accuracy, ROC and Kappa, respectively with 7 features. The PSO-based feature selection technique highest performance rate is associated with Random Forest classifier with 95.1%, 97% and 84.5% for Accuracy, ROC and Kappa, respectively with 6 features.

Table 6

Performance measures of selected features after SRS

Classifier	Performance index	IGR	SU	GA	PSO
MLP	Accuracy	0.888	0.915	0.861	0.855
	ROC	0.842	0.873	0.839	0.837
	Kappa	0.641	0.716	0.537	0.521
Simple Logistic	Accuracy	0.845	0.857	0.840	0.841
	ROC	0.861	0.874	0.838	0.839
	Kappa	0.476	0.491	0.456	0.461
Random Forest	Accuracy	0.979	0.995	0.962	0.951
	ROC	0.987	0.997	0.972	0.970
	Kappa	0.933	0.982	0.881	0.845
C4.5	Accuracy	0.936	0.984	0.925	0.906
	ROC	0.932	0.986	0.919	0.900
	Kappa	0.797	0.947	0.761	0.698
FURIA	Accuracy	0.962	0.961	0.911	0.890
	ROC	0.970	0.969	0.913	0.848
	Kappa	0.8752	0.8718	0.7046	0.625

Figure 4 visualizes the feature selection agreements between the IGR, SU, GA and PSO models. The Venn diagram shows the suggested models share student's gender, home to school travel time, going out with friends, current health status, number of school absences and final grade, in which all was obtained by the PSO model.

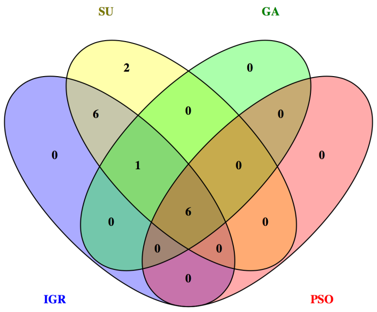

Figure 4 PSO feature selection agreement in the student alcohol consumption dataset

PSO is a well-known optimization method that has a strong search capability and usually used for fine-tuning of the features space. Our proposed technique based on PSO succeeded in significantly improving the classification performance with a limited number of features compared to other techniques. The suggested PSO selection-based features demonstrated accuracies between 84.1% and 95.1% in various DM model and this is a quite high performance for predicting student performance [29]. Therefore, deploying PSO in feature selection clearly helped in reducing the size of dataset from 33 to only 6 attributes. It is worth noting that as the number of features has reduced, less storage space and classification complexity is further required. Moreover, the results demonstrated that these features are adequate to represent the dataset's class information. The outcomes from the suggested feature selection techniques show better results compared to datasets which are not pre-processed and also when these attribute selection techniques are used independently. As can be seen from above results, the proposed technique based on PSO has produced very promising results on the classification of multi-class dataset in predicting the student alcohol consumption.

As we can comprehend from the data graph and table there is a significant gender differences in the drinking habits. Comparing to men, women are more likely to be responsive to health concerns and are less likely to engage in risky health behaviours [10,11]. Commonly, men smoke and drink more than women in different societies and cultures, and women have a higher expectation of self-control than do men [12,13]. That can lead the other features such as the free time with friends, high travel time between school and home and the number of school absences and eventually the poor academic performance. Our study shows that, drinking is the product of many factors working together. This suggests that the educational professionals can consider these features for further analysis in future.

5. Conclusion

Underage drinking or adolescent alcohol misuse is a major public health concern. The proposed machine learning approach could improve the accuracy performance and achieve promising results in identifying risk or protective factors for high-risk drinking that can be used to help detect and address the early developmental signs of alcohol abuse and dependence within adolescent students. The experiment results have shown that the PSO helped in reducing the feature space, whereas adjusting the original data with simple random sampling helped in increasing the region area of the minority class in favour of handling the existing imbalanced data property.

Acknowledgement

The authors would like to kindly appreciate and gratefully acknowledge, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] and Mr. Paulo Cortez [11] for obtaining the student performance dataset.

References

1. World Health Organization Management of Substance Abuse Team. Global Status Report on Alcohol and Health. World Health Organization, Geneva, Switzerland; 2011:85.

2. GRANT, B.F. Estimates of U.S. children exposed to alcohol abuse and dependence in the family. American Journal of Public Health 90(1):112–115, 2000.

3 Aertgeerts B, Buntinx F. The relation between alcohol abuse or dependence and academic performance in first-year college students. J Adolesc Health. 2002; 31:223–5.

4. Berkowitz AD, Perkins HW. Problem drinking among college students: A review of recent research. J Am Coll Health. 1986;35:21–8.

5. Martinez JA, Sher KJ, Wood PK. Is heavy drinking really associated with attrition from college? The alcohol-attrition

paradox. Psychol Addict Behav. 2008;22:450–6.

6. Crutzen, R., P.J. Giabbanelli, A. Jander, L. Mercken and H. de Vries, 2015. Identifying binge drinkers based on parenting dimensions and alcohol-specific parenting practices: Building classifiers on adolescent-parent paired data. BMC Public Health, 15(1): 747.

7. Montaño, J.J., E. Gervilla, B. Cajal and A. Palmer, 2014. Data mining classification techniques: An application to tobacco consumption in teenagers. An. Psicol., 30(2): 633-641.

8. Pang, R., A. Baretto, H. Kautz and J. Luo, 2015. Monitoring adolescent alcohol use via multimodal analysis in social multimedia. Proceeding of the IEEE International Conference on Big Data (Big Data), pp: 1509-1518.

9. Bi, J., J. Sun, Y. Wu, H. Tennen and S. Armeli, 2013. A machine learning approach to college drinking prediction and risk factor identification. ACM T. Intell. Syst. Technol., 4(4).

10. Zuba, M., J. Gilbert, Y. Wu, J. Bi, H. Tennen and S. Armeli, 2012. 1-norm support vector machine for college drinking risk factor identification. Proceeding of the 2nd ACM SIGHIT International Health Informatics Symposium, pp: 651-660.

11. Cortez, P. and Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In Proceedings of 5th Future Business Technology Conference. pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

12. Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L.A. (2006). Feature Extraction, Foundations and Applications, Springer, Berlin,

13. Kennedy, J. and Eberhart, R. (2001). Swarm intelligence. Morgan Kaufmann.

14. Kennedy, J. and Eberhart, R. (1997). A discrete binary version of the particle swarm algorithm. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol.5, pp. 4104–4108.

15. Quinlan J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

16. Fayyad, U., and Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (pp. 1022–1027). Morgan Kaufmann.

17. J.H. Holland ,Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI (1975).

18. Randy, H., and Haupt, S., 1998. Practical Genetic Algorithms, John Wiley and Sons.

19. David E. Goldberg (1989). Genetic algorithms in search, optimization and machine learning. Addison-Wesley.

20. Hall, Mark A. Correlation-Based Feature Selection for Machine Learning, 1999.

21. G. Weiss and F. Provost, "Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction," J. Artificial Intelligence Research, vol.19,315-354,2003.

22. Park, B.-H., Ostrouchov, G., Samatova, N.F., Geist, A.: Reservoir-based random sampling with replacement from data stream. In: SDM 2004 , 492-496, (2004)

23. Mitra SK and Pathak PK. The nature of simple random sampling. Ann. Statist., 1984, 12:1536-1542.

24. Pagnotta, F. and H.M. Amran, 2016. Using data mining to predict secondary school student alcohol consumption. Department of Computer Science, University of Camerino.

25. Fawcett, T. and Provost, F. (1996). Combining data mining and machine learning for effective user profiling. In Proceedings of KDD-96, 8-13. Menlo Park, CA: AAAI Press.

26. Fleiss, J.L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33: 613–619.

27. Devijver P.A., and Kittler J. (1982). Pattern Recognition: A Statistical Approach. London, GB: Prentice-Hall.

28. Gupta G.K. (2006). Introduction to Data Mining with Case Studies. Prentice-Hall of India.

29. Liao S., Chu, P. and Hsiao, P.(2012). Data mining techniques and applications. A decade review from 2000 to 2011, Expert Systems with Applications 39.